Session 7: Linked Repositories – Theme of the Day

Moderator: Barbara Butler

Using Linked Open Data and Semantic Integration to Search Across Repositories

Lisa Raymond, Audrey Mickle, and GeoLink Project Group

Woods Hole Oceanographic Institution
MS#8, Woods Hole, MA 02543

Abstract:
The MBLWHOI Library is a partner in the GeoLink project, an NSF EarthCube Building Block, applying semantic technologies to enable knowledge discovery, sharing and integration. GeoLink is testing ontology design patterns that link together the MBLWHOI Library Institutional Repository (IR) Woods Hole Open Access Server (WHOAS); data repositories, including Rolling Deck to Repository (R2R), Biological and Chemical Oceanography Data Management Office (BCO-DMO), Integrated Earth Data Applications (IEDA), Long-Term Ecological Research Network (LTER), DataONE and the International Ocean Discovery Program (IODP); the National Science Foundation (NSF) funded awards; and American Geophysical Union (AGU) conference presentations.

Keywords: Linked Open Data, Semantic Web, repositories.

The Library is collaborating with scientific users, data managers, DSpace engineers, experts in ontology design patterns, and user interface developers to make WHOAS, a DSpace repository, available as linked open data. The goal is to enable searching across participating repositories without needing to change the way information providers are managing their content. The tools developed for DSpace will be made freely available to the community of users. There are 257 registered DSpace repositories in the United States and over 1700 worldwide.
Outcomes include: Integration of DSpace with OpenRDF Sesame triple store to provide SPARQL endpoint for the storage and query of RDF representation of DSpace resources, mapping of DSpace resources to Geolink ontology, and DSpace "data" add on to provide resolvable linked open data representation of DSpace resources.

Linked Open Data (LOD, Berners-Lee, 2006) is an approach to publishing and linking content online using methods and protocols known as the Semantic Web (Hitzler et al, 2010). The methodology has four basic characteristics: 1) Use of unique identifiers (URIs) to name (identify) things; 2) Use of HTTP URIs so that things can be looked up ("dereferenced") on the Web; 3) Provide useful information about what a name identifies when it is looked up using open standards such as the Resource Description Framework (RDF) language (Berners-Lee, 2006, Schreiber and Raimond, 2014) when identifiers are dereferenced; and 4) Refer to other things using their HTTP URI when publishing on the Web to support further discovery. LOD alone can

succeed in opening up access to information, but it does not make data readily reusable for scientific purposes (Janowicz and Hitzler, 2012). LOD capabilities for scientific use and data discovery are greatly improved when further supported by standard vocabularies such as W3C DCAT (Mali and Erikson, 2014) and PROV (Lebo et al, 2013) that describe datasets and provenance, along with the OGC GeoSPARQL language to perform queries against RDF data (Battle et al 2012). Concepts such as publications, researchers, expeditions, etc. should be modeled in a modular and consistent way using ontology design patterns (Janowicz and Hitzler, 2012) to maximize reusability. Modularity allows for use and re-use of scenarios by different repositories. The MBLWHOI Library is a partner in the GeoLink project, an NSF EarthCube Building Block, applying these semantic technologies to enable knowledge discovery, sharing and integration.

The current project is an extension of a narrower effort called OceanLink that had five data sources: MBLWHOI Library Institutional Repository (IR) Woods Hole Open Access Server (WHOAS); Rolling Deck to Repository (R2R); Biological and Chemical Oceanography Data Management Office (BCO-DMO); the National Science Foundation (NSF) funded awards; and American Geophysical Union (AGU) conference presentations. All of these contain different kinds of data, are built on different platforms, and are organized in different way, but we created the ontology design patterns that enabled the connections. The Library collaborated with the open-source community of developers, including @mire, Inc., a Registered Duraspace service provider, to make our DSpace repository LOD enabled. The result of this first phase of work was an add-on for DSpace that is available on GitHub https://github.com/dspace-oceanlink/DSpace/tree/oceanlink-4_x/dspace-lod

GeoLink attempts to go further, addressing disparate vocabularies and heterogeneity issues regarding representation of information across the geosciences. We are using a conceptual modeling approach without needing to change the way information providers are managing their content. GeoLink is testing ontology design patterns that link together the MBLWHOI Library Institutional Repository (IR) Woods Hole Open Access Server (WHOAS); data repositories, including Rolling Deck to Repository (R2R), Biological and Chemical Oceanography Data Management Office (BCO-DMO), Integrated Earth Data Applications (IEDA), Long-Term Ecological Research Network (LTER), DataONE and the International Ocean Discovery Program (IODP); the National Science Foundation (NSF) funded awards; and American Geophysical Union (AGU) conference presentations.

The MBLWHOI Library is continuing to work with @mire on DSpace enhancements. Editable administrative authority control, which provides a local authority storage solution for DSpace, has been deployed to WHOAS. Functionality includes the ability for administrative users to define and assign authority control to DSpace metadata fields for use during submission and deposit. This solution includes support for SPARQL endpoints as authoritative sources. It also provides a mechanism to define and configure queries against known endpoints, allowing retrieval of specific resources associated with the metadata fields in DSpace records. Additional edits were done to update this work to the latest major DSpace version release, DSpace 5, and to map DSpace 5 RDF output to the latest GeoLink ontology release.

GeoLink has established resource discovery across geoscience repositories using LOD and semantic integration while respecting and preserving the heterogeneous landscape of data providers. The Library has a long history of partnering with groups interested in increasing access to research and data across the geosciences. Through the GeoLink project we have been able to make significant progress in that mission, linking our openly available theses, articles, and data with primary and axillary data and metadata across organizations and disciplines without changing the structure of our records or databases.

References

Battle, R. and Kolas, D. 2012. Enabling the Geospatial Semantic Web with Parliament and GeoSPARQL. *Semantic Web* 3(4), 355-370.

Berners-Lee, T. 2006. Linked Data. http://www.w3.org/DesignIssues/LinkedData.html, June 27, 2006, edited June 18, 2009.

Hitzler, P., Krötzsch, M., Rudolph, S. 2010. *Foundations of Semantic Web Technologies.* Boca Raton: CRC/ Chapman and Hall. 456 p.

Janowicz, K., Hitzler, P. 2012. The Digital Earth as knowledge engine. *Semantic Web* 3(3), 213-221.

Lebo, T., Sahoo, S., McGuinness, D. (2013). PROV-O: The PROV Ontology. W3C Recommendation 30 April 2013 http://www.w3.org/TR/prov-o/.

Maali, F., Erikson, J., 2014. Data Catalog Vocabulary (DCAT)W3C Recommendation 16 January 2014 http://www.w3.org/TR/vocab-dcat/.

Schreiber, G., Raimond, Y. 2014. RDF 1.1 Primer 25 February 2014. W3C Working Group Note http://www.w3.org/TR/rdf11-concepts/.